Lead Score Case Study Summary

By:
Thulasiram Saravanan
and Swapnil Kudale

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.
The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

Step 1: Reading and Understanding Data:

Read and inspected the data.

Step 2: Data Cleaning

a. First step is to clean the data and the choice was to drop the columns having more than 40% missing values.
b. Updated values of Nan in City to the categories without affecting the current percentage.
c. Updated the Specialization, Tags Nan values to Others and Unknown.
d. Dropped the columns because they don't have even distribution,
    i. 'Prospect ID',
    ii. 'Lead Number',
    iii. 'What matters most to you in choosing a course',
    iv. 'Country'
e. Filling the missing values in numerical variables by median.
f. Replacing the null values in categorical columns having few null values only using mode.
g. One column was having identical label in different cases (first letter small and capital respectively). Fixed this issue by converting the label with first letter in small case to upper case.
h. The outlier analysis of the numerical variables is performed and have removed the outliers and retained 0.995 quantile value of Data frame.

Step 3: Exploratory data analysis

1. Visualized various plots of univariant, Bi variant and Multivariant analysis and understood the form and spread of data and its features.
2. By observing the plots and data correlation that the problem can be solved using Logistic regression as there is no direct or liner dependency between any of the variables and the target variables.

Step4: Dummy Variables Creation:

Created dummy variables for the categorical variables. Removed all the repeated and redundant variables.

Step 5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 80-20% values using Stratified function to maintain the imbalance ratio of the target variable.

Step 6: Scaling the dataset

a. The Training data numerical features are Standard scaled and fit
b. The testing data numerical features are scaled using standard scaler but not fit.

Step 7: Model Building

a. Using RFE, top 18 features are selected.
b. Added constant variable to the X_train column.
c. By recursively training the model each time using logistic regression eliminating variables under below conditions final model is derived,
    i. Each time the model is trained, each of the variables P value and VIF are checked.
    ii. The P value must be less than 0.05 and VIF values should be as low as possible within a certain range.
    iii. Firstly, starting with High VIF values followed by high P values drop for each trial one at a time and the final list of variables is 15.

Step 8: Model Evaluation

a. The built model is evaluated by predicting the train and the test data and deriving the predicted value.
b. ROC plot is laid to find the optimal cut off point between the specificity, sensitivity and accuracy, found to be 0.3.
c. By Precision- Recall trade off, the optimal cut off as 0.37. But considering to increase the probability to target variable being true, we set the optimal cut off as 0.27.
d. The metrics obtained for Train data are
    i. Precision = 84.74%
    ii. Recall = 90.41 %
    iii. Accuracy = 90%
e. The metrics obtained for Test data are
    i. Precision = 84.68%
    ii. Recall = 90.79 %
    iii. Accuracy = 90.13%.

Conclusion:

- **To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and Google leads and generate more leads from reference and welingak website.**
- **Lead conversion rate, can be improved by focusing more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.**
- **Though Google is the highest source to get leads, the lead conversion through Google is low comparatively.**
- **Focus on Working Professional which has high conversion.**
- **Website should be made more engaging to make leads spend more time.**
- **Improve the Olark Chat service since this is affecting the conversion negatively.**